1 General information

GATE Noun/Verb Phrase Chunker

Gate group[1], Department of Computer Science, University of Sheffield, UK

Status: ready to use

Languages covered: English

2 Implemented NLP services

The chunker produces text annotated at phrase level in xml format.
For producing this annotation output it depends on the linguistic preprocessing of the text input (for required text format see below) with domain- and application-independent techniques.

- Tokenization: the tokeniser splits text into simple tokens, such as numbers, punctuation, symbols, and words of different types (e.g. with an initial capital, all upper case, etc.).
- The sentence splitter segments the text into sentences.
- The part-of-speech tagger adds morphosyntactic information to each token.

3 Language resource standards

No language resource standards are required for input.

Output:
The pos tags are Penn Treebank and MAF compliant.
The chunking is SYNAF compliant.

4 Linguistic data encoding

The service does not assume any data categories as input.
It will run over a text corpus containing documents from a large variety of formats:
plain text
HTML

[1] http://www.gate.ac.uk

SGML
XML
RTF
PDF (not all)
Microsoft Word (not all)